Geographically constrained information retrieval

Andogah, Geoffrey

Chapter 7

Relevance ranking

[Parts of of this chapter are published in Andogah and Bouma (2008).]

In the context of information retrieval, *relevance* denotes how well a retrieved set of documents (or a single document) meets the information need of the user. Relevance most commonly refers to topical relevance or aboutness, i.e. to what extent the topic of a result matches the topic of the query or information need. Relevance can also be interpreted more broadly, referring to generally how "good" a retrieved result is with regard to the information need. The latter definition of relevance, sometimes referred to as user relevance, encompasses topical relevance and possibly other concerns of the user such as timeliness, authority or novelty of the result.[1] *Relevance ranking* is the task of ordering the retrieved set of documents by relevance to the user's information needs so that the most relevant documents are pushed to the top of the ranked result list.

The research objective addressed in this chapter is – How well can geographical scope and feature type information be integrated into the document ranking procedure to prioritize documents by geography? The chapter describes two types of relevance ranking schemes which exploit geographical scopes and feature types in documents and search queries to rank documents by geography. The *scope-based* metric is used to rank documents for queries which are resolvable to at least one scope. On the other hand, the *type-based* metric is used to rank documents when a query mentions only the geographical subjects, therefore, it is not resolvable to a scope. The scores of the non-geographic component and the geographic components are combined through linear interpolation and through weighted harmonic means.

[1][14 March, 2010]: `http://en.wikipedia.org/wiki/Relevance_(information_retrieval)`

The proposed relevance measures and weighting schemes are evaluated on the GeoCLEF 2007 dataset with an encouraging performance improvement over the standard IR performance. The best performance is achieved when the importance of non-geographic relevance scores outweighs the importance of geographic relevance scores.

7.1 Non-geographic relevance measure

The Apache Lucene search engine library is used to perform non-geographic search. Lucene's default relevance measure is derived from the vector space model (VSM) (Salton, 1989). The Lucene relevance score formula combines several factors to determine the score of a document for a given query (Gospodnetic and Hatcher, 2005):

$$NonSim(q,d) = \sum_{t \in q} tf(t \in d) \cdot idf(t) \cdot bst \cdot lN(t.field \in d) \qquad (7.1)$$

where, $\cdot$ operator stands for multiplication, $tf(t \in d)$ is the term frequency factor for term t in document d, $idf(t)$ is the inverse document frequency of term t in the document collection, bst is the field boost set during indexing and $lN(t.field \in d)$ is the normalization value of a field given the number of terms in the field (see Gospodnetic and Hatcher (2005) for more details). The purpose of boosting is to indicate how important a given term is relative to other terms in the document.

7.2 Scope-based relevance measure

The scope-based relevance measure (SBRM) uses geographical scopes assigned to queries and documents to rank documents according to query geographic restrictions similar to schemes explored in Andrade and Silva (2006). The geographical scope resolver (Andogah et al., 2008) assigns multiple scopes to a document. The assigned scopes are ranked according to their relevance score from the most relevant to the least relevant per document.

The scopes are limited to six categories: continent scope, continent-directional scope (e.g. western Europe), country scope, country-directional scope (e.g. north Netherlands), province[2] scope, and province-directional scope (e.g. south-east California).

[2]A province is the first order administrative division of a country.

The scope-based relevance measure is defined as:

$$ScopeSim(q,d) = \sum_{s} \sqrt{wt_{(q,s)}} \times log(1 + wt_{(d,s)}) \tag{7.2}$$

where $wt_{(q,s)}$ is the weight assigned to scope s in query q by the scope resolver, and $wt_{(d,s)}$ is the weight assigned to scope s in document d by the scope resolver. The SBRM is applicable to queries with explicit geographical scopes. The scopes are assigned using the geographical scope resolver described in Chapter 4.

For instance, consider query q_1 with geographical scopes s_1, s_2, s_3 with the following scores:

(1) *Query:*
q_1: s_1 with score *4*, s_2 with score *9* and s_3 with score *16*

Further more, consider three documents d_1, d_2, d_3 with the following geographical scopes:

(2) *Documents:*

a. d_1: s_1 with score *9*, s_2 with score *4*, s_3 with score *2*
b. d_2: s_1 with score *9*, s_2 with score *4*
c. d_3: s_4 with score *9*, s_2 with score *16*, s_3 with score *1*

Using Equation 7.2 document relevance scores are computed as follows:

$$ScopeSim(q_1,d_1) = \sqrt{4} \times log(10) + \sqrt{9} \times log(5) + \sqrt{16} \times log(3) = 6.0$$
$$ScopeSim(q_1,d_2) = \sqrt{4} \times log(10) + \sqrt{9} \times log(5) = 4.1$$
$$ScopeSim(q_1,d_3) = \sqrt{9} \times log(17) + \sqrt{16} \times log(2) = 4.9$$

The documents are geographically ranked from the most relevant to the least relevant as - d_1, d_3, d_2 with scores of 6.0, 4.9 and 4.1 respectively.

7.3 Type-based relevance measure

The type based relevance measure (TBRM) utilizes the geographical feature class and type defined in a database of geographic features to compute a document's relevance to a query. The measure ranks documents by query feature type restriction. The feature class and type as defined in the Geonames.org[3] database are used to implement the type-based relevance measure.

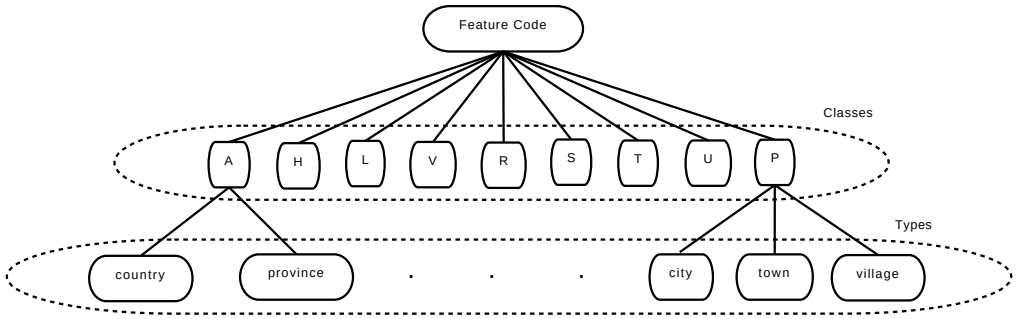

Figure 7.1: Sample Geonames.org feature code hierarchy.

Figure 7.1 shows the structure of the Geonames.org [4] feature grouping hierarchy, where, A is administrative unit, H is hydrographic, L is locality or area, P is populated place, R is road or railroad, S is spot, T is hypsographic, U is undersea and V is vegetation. The type based relevance measure is defined as:

$$TypeSim(q,d) = \frac{1.0}{\sqrt{1 + \frac{N_{qFClass} - N_{qFType}}{N_{qFClass}}}} \tag{7.3}$$

where; $N_{qFClass}$ is the number of occurrences of the required query feature class in the document, and N_{qFType} is the number of occurrences of the required query feature type in the document. The maximum value of 1.0 is reached for Equation 7.3 when the number of $N_{qFClass}$ is equal to N_{qFType}. This happens when all features of class $FClass$ are of type $FType$. The TBRM is applicable to queries that mention geographical subjects or types without the mention of place names, e.g., *lakes with monsters*. The use of Equation 7.3 illustrated with GeoCLEF 2007 (Mandl et al., 2008) topic *10.2452/56-GC: Lakes with monsters*. Figure 7.2 shows geographic feature types in three relevance documents to topic *10.2452/56-GC*.

The query feature type *Lake (LK)* belongs to class H (i.e., hydrographic features such as river, stream, lake, bay, etc.). Each retrieved document is queried for class H and feature type LK. The number of occurrence of H (i.e. $N_{qFClass}$) and LK (i.e. N_{qFType}) in the document is used to compute the document's geographic relevance according to Equation 7.3. The documents in Figure 7.2 are geographically ranked as: *LA071094-0288* , *LA090394-0008* and *GH950721-000028* with scores of *1.0*, *0.89* and *0.77* respectively.

[3] `http://www.geonames.org/export/codes.html`
[4] `http://www.geonames.org`

```
<Document did="LA071094-0288">
  <GT name="village" tf="1" type="PPL" class="P" />
  <GT name="lake" tf="1" type="LK" class="H" />
  <GT name="town" tf="1" type="PPL" class="P" />
</Document>

<Document did="GH950721-000028">
  <GT name="sea" tf="4" type="SEA" class="H" />
  <GT name="lake" tf="2" type="LK" class="H" />
</Document>

<Document did="LA090394-0008">
  <GT name="sea" tf="1" type="SEA" class="H" />
  <GT name="lake" tf="3" type="LK" class="H" />
  <GT name="city" tf="2" type="PPL" class="P" />
</Document>
```

Figure 7.2: Sample documents with geographic feature classes and types.

7.4 Relevance measure unification

This section describes attempts that have been proposed to combine the non-geographic relevance measures and the geographic relevance measures to unified relevance measures.

7.4.1 Linear interpolated combination

The linear interpolated combination (LIC) is derived as:

$$LIC(q,d) = \lambda_T \, NonSim(q,d) + \lambda_G \, GeoSim(q,d) \tag{7.4}$$

$$\lambda_T + \lambda_G = 1 \tag{7.5}$$

where; λ_T is the non-geographic interpolation factor (NIF) and λ_G is the geographic interpolation factor (GIF). The non-geographic and geographic scores are normalized to $[0, 1]$ before linearly combining the ranked lists. The $GeoSim(q, d)$ in Equation 7.4 is replaced by either Equation 7.2 or Equation 7.3 depending on the nature of the query.

7.4.2 Weighted harmonic mean combination

The weighted harmonic mean (WHM) combination borrows from the classic precision and recall combination formula, the *F-measure* (see Van Rijsbergen, 1979, Chapter 7) commonly used to measure performance of information retrieval (IR) systems. The motivation is to determine the importance of non-geographic relevance relative to geographic relevance, and then use the insight to rank documents by both non-geographic and geographic relevance. The weighted harmonic mean (WHM) combination is defined as:

$$WHM(q,d) = \frac{(1+\beta) \times GeoSim(q,d) \times NonSim(q,d)}{\beta \times GeoSim(q,d) + NonSim(q,d)} \tag{7.6}$$

where; β indicates the importance attached to either $GeoSim(q, d)$ or $NonSim(q, d)$ in the unification. The following special cases are derived as a consequence of harmonic mean combination:

1. if $\beta = 1$, equal importance is attached to both non-geographic and geographic relevance.

2. if $\beta = 0$, no importance is attached to non-geographic relevance.

3. if $\beta = \infty$, no importance is attached to geographic relevance.

The interesting feature of this combination is that an optimal value of β where the best performance is achieved can be spotted. The $GeoSim(q, d)$ in Equation 7.6 is replaced by either Equation 7.2 or Equation 7.3 depending on the nature of the query.

7.4.3 Extended harmonic mean combination

The extended harmonic mean (EHM) combination linearly adds the non-geographic relevance measure (see Eq. 7.1) to the weighted harmonic mean (WHM) combination (see Eq. 7.6) as follows:

$$EHM(q,d) = NonSim(q,d) + \frac{(1+\beta) \times GeoSim(q,d) \times NonSim(q,d)}{\beta \times GeoSim(q,d) + NonSim(q,d)} \tag{7.7}$$

The $GeoSim(q, d)$ in Equation 7.7 is replaced by either Equation 7.2 or Equation 7.3 depending on the nature of the query.

7.5 Evaluation

The proposed relevance measure and weighting schemes are evaluated on GeoCLEF 2007 (Mandl et al., 2008) dataset. Gey et al. (2007) categorized geographic topics into eight groups according to the way they depend on a place (e.g., Netherlands, Texas, etc), geographic subject (e.g., city, river, etc.) or geographic relation (e.g., north Groningen, western Europe, etc.). The ranking parameters in the formula for the experiment are tuned on the GeoCLEF 2006 dataset (i.e., the topics and document collection). The ultimate purpose of the experiment is to compare the proposed relevance ranking schemes against the default search engine relevance ranking. Therefore, efforts were made to construct high quality queries to run against the document collections.

In this experiment two groups of topics are distinguished:

1. GROUP1: Topics which explicitly mention places of interest by name, and are resolvable to geographical scopes. Topics which lack sufficient geographical information or provide ambiguous geographic information are geographically expanded. To generated high quality queries, geographical expansion is done manually. For instance, GeoCLEF 2007 topic *10.2452/65-GC: Free elections in Africa* is manually expanded by adding the names of African countries and their capitals. The GeoCLEF 2007 topics geographically expanded include: 10.2452/51-GC, 10.2452/59-GC, 10.2452/60-GC, 10.2452/61-GC, 10.2452/63-GC, 10.2452/65-GC, 10.2452/66-GC, 10.2452/70-GC.

Example GROUP1 Topic	
Topic num	10.2452/65-GC
Topic title	Free elections in Africa
Geo-expansion	Add names of African countries and their capitals
Query	Formulated by content of topic title-desc-narr tags
Geo-relevance	Scope based measure (see Eq.7.2)
Example GROUP2 Topic	
Topic num	10.2452/68-GC
Topic title	Rivers with floods
Geo-expansion	–
Query	Formulated by content of topic title-desc-narr tags
Geo-relevance	Type based measure (see Eq.7.3)

Table 7.1: Example topic grouping and query formulation

2. GROUP2: Topics which do not mention places, but mention a geographic subject of interest (i.e., 10.2452/56-GC, 10.2452/67-GC, 10.2452/68-GC, 10.2452/72-GC). Topics 10.2452/56-GC, 10.2452/68-GC, 10.2452/72-GC are characterized as *geographical subjects with non-geographic restriction* (Gey et al., 2007). Topic 10.2452/67-GC is more complex. Resolving the geographic scope of such topics to a specific place is non-trivial. The most reasonable scope for these topics is a geographic subject scope: *lake, river, beach, city,* etc.

Table 7.1 shows GeoCLEF 2007 topics[5] depicting the topic grouping, geographic expansion, query formulation and geographic relevance measure used. GROUP1 topics are ranked with the scope-based relevance ranking scheme because they either inferred or explicitly mentioned names of places, and therefore were assigned geographical scopes based on this. On the other hand, GROUP2 topics are ranked with the type-based relevance ranking scheme because they only mentioned geographical subjects of interest without mentioning names of places, therefore, no geographical scope was assigned to them.

7.5.1 Harmonic mean vs. linear interpolated combination

This sub-section compares harmonic mean combination (see Equation 7.6) retrieval performance against linear interpolated combination (see Equa-

[5]See Section B.3 in Appendix B for a complete list of GeoCLEF 2007 topic titles.

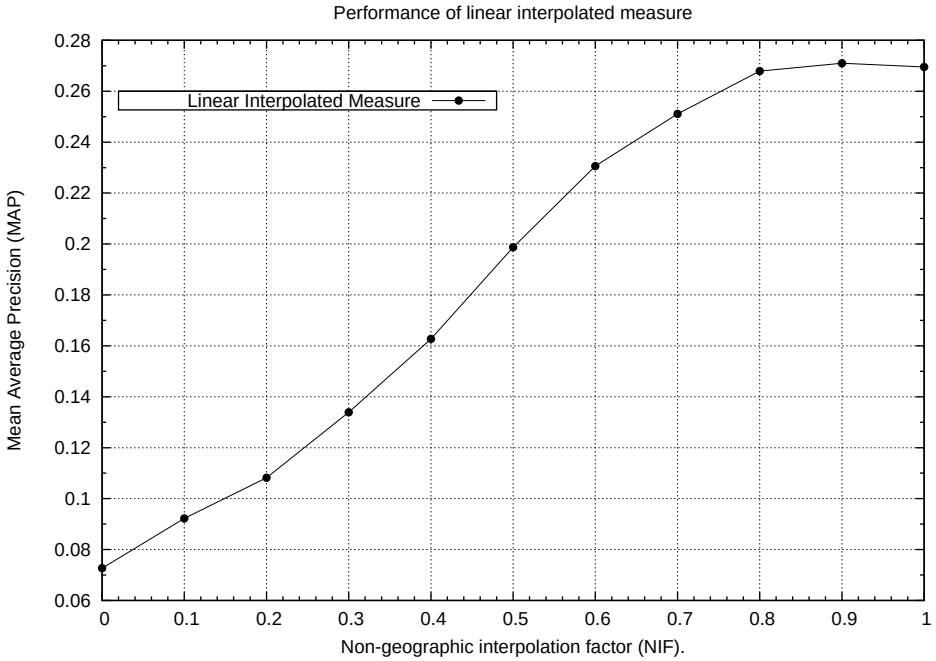

Figure 7.3: Variation of MAP as a factor of NIF λ_T.

tion 7.4).

Figure 7.3 and Figure 7.4 shows result of the experiment with GeoCLEF 2007 dataset, and the following observations are made:

1. The system performance is worst at $\lambda_T = 0$ and $\beta = 0$ which represents pure geographic retrieval.

2. The system performance is below average at $\lambda_T = 0.5$ and $\beta = 1$ which give equal importance to geographic retrieval and non-geographic retrieval in comparison to pure non-geographic retrieval at $\lambda_T = 1.0$ and $\beta = \infty$.

3. The best system performance is observed at $\lambda_T = 0.9$ with a MAP of 0.2710 (Fig. 7.3) and $\beta \geq 50$ with a MAP of 0.2749 (Fig. 7.4).

There is no significant difference between harmonic mean-based measure (i.e., with MAP score of 0.2749) and the linear interpolated measure (i.e., with MAP score of 0.2710). However, the harmonic mean-based measure achieves slightly better performance improvement of *2.0%* over the default Lucene setting (i.e., with a MAP score of 0.2695). The best performance is achieved when the importance of non-geographic relevance outweighs the importance of geographic relevance. Buscaldi and Rosso (2008) reported an

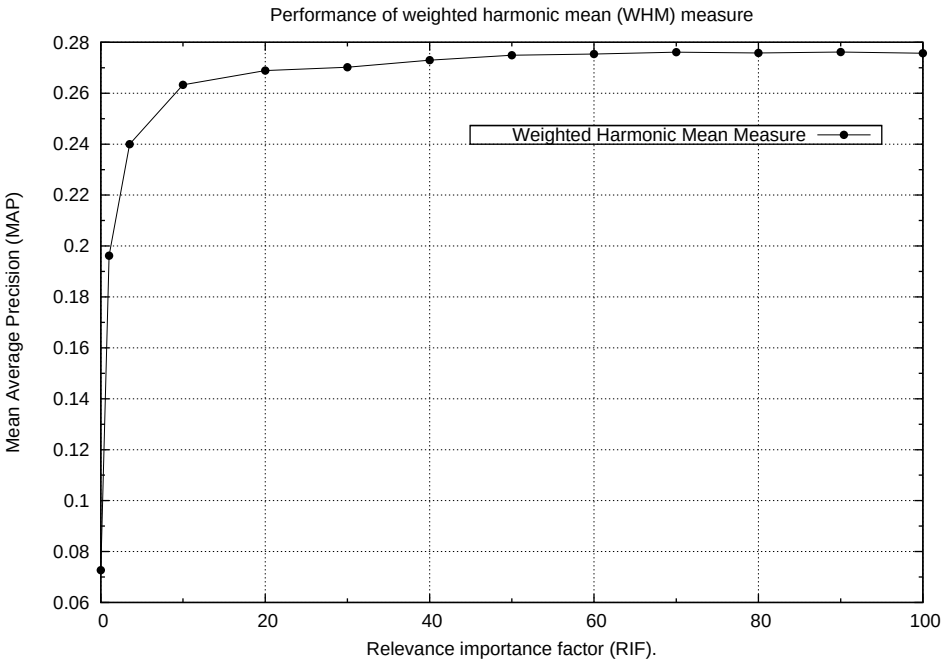

Figure 7.4: Variation of MAP as a factor of RIF β.

improvement when geographic terms in the query are weighed half or less than the weight of non-geographic terms, which is in agreement with our observation. The difference is that geographical scopes are used here instead of geographic terms.

7.5.2 Extended harmonic mean combination

For the extended harmonic mean (EHM) evaluation, two variant of Equation. 7.2 are implemented:

$$ScopeSimB(q,d) = \sum_{s} \sqrt{wt_{(q,s)}} \times \sqrt{wt_{(d,s)}} \tag{7.8}$$

and

$$ScopeSimC(q,d) = \sum_{s} \sqrt{wt_{(d,s)}} \times log(1 + wt_{(q,s)}) \tag{7.9}$$

The graphs $ScopeSimA$, $ScopeSimB$ and $ScopeSimC$ in Figure 7.5 show the performance of EHM with $GeoSim(q,d)$ in Equation 7.7 replaced by Equations 7.2, 7.8 and 7.9 respectively. In Figure 7.5, it can be seen that the $GeoSim(q,d)$ formula in Equation 7.2 gives the best performance for β with the value greater than *3.5* achieving the MAP score of *0.2935* which

Rank	Participant	MAP
1^{st}	catalunga	28.50%
2^{nd}	cheshire	26.42%
3^{rd}	valencia	26.36%
4^{th}	groningen	25.15%
5^{th}	csusm	21.32%
	EHM	29.35%

Table 7.2: Comparison to GeoCLEF 2007 participants.

presents a *8.9%* improvement over the default Lucene score (with a MAP score of *0.2695*).

Figure 7.6 shows topic performance with extended harmonic mean (EHM) using formula in Equation 7.7 against default Lucene measure (see Equation 7.1). For topics 10.2452/56-GC, 10.2452/67-GC, 10.2452/68-GC and 10.2452/72-GC, $GeoSim(q, d)$ in Equation 7.7 is replaced with type based relevance measure (see Equation 7.3). For the rest of the topics, $GeoSim(q, d)$ in Equation 7.7 is replaced with scope-based relevance measure in Equation 7.2.

Table 7.2 shows the best five entries in GeoCLEF 2007 campaign (Mandl et al., 2007) where there were eleven competing teams in total. The *groningen* row is the performance of our system in the campaign (Andogah and Bouma, 2007). EHM shows the performance of relevance ranking formula in Equation 7.7. The EHM performs slightly better than the best entry *catalunya* by a margin of 2.98%.

With properly balanced contributions of the non-geographic and geographic scores to a combined score (see Equation 7.7), an improvement can be achieved. The optimal integration of geographical scope information to improve search engine relevance ranking is still an open question that needs further investigation.

This chapter demonstrated the following:

1. Geographical information in user query and document collection can be exploited to improve the performance of standard search engine systems when more importance is attached to non-geographic relevance than geographic relevance.

2. Geographical scope and type information can be used to construct relevance measures to rank documents by geography.

3. A weighted harmonic mean combination of non-geographic and geographic relevance is a better option than linear interpolation.

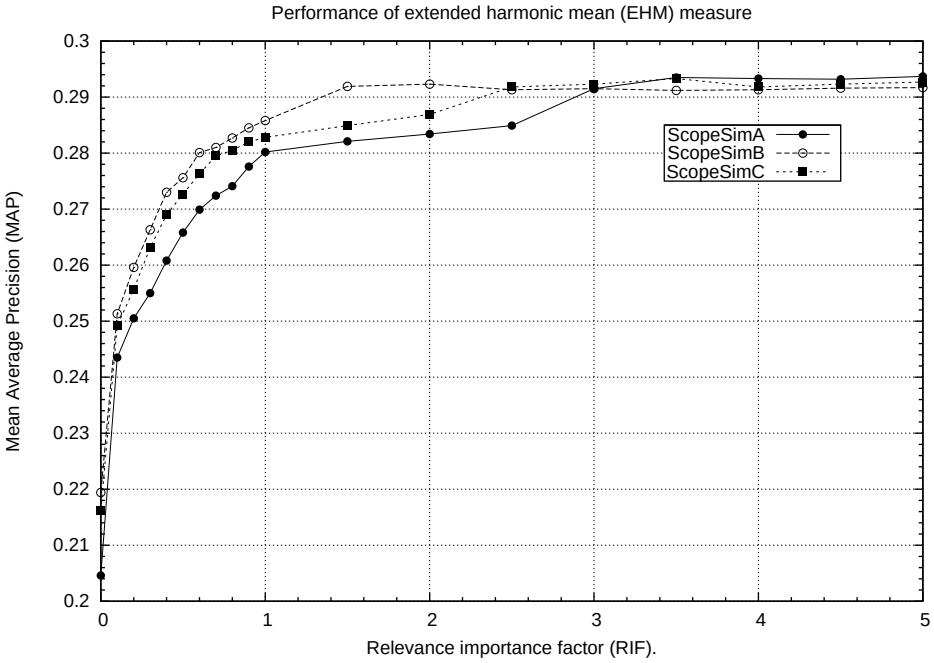

Figure 7.5: Variation of MAP as a function of RIF β.

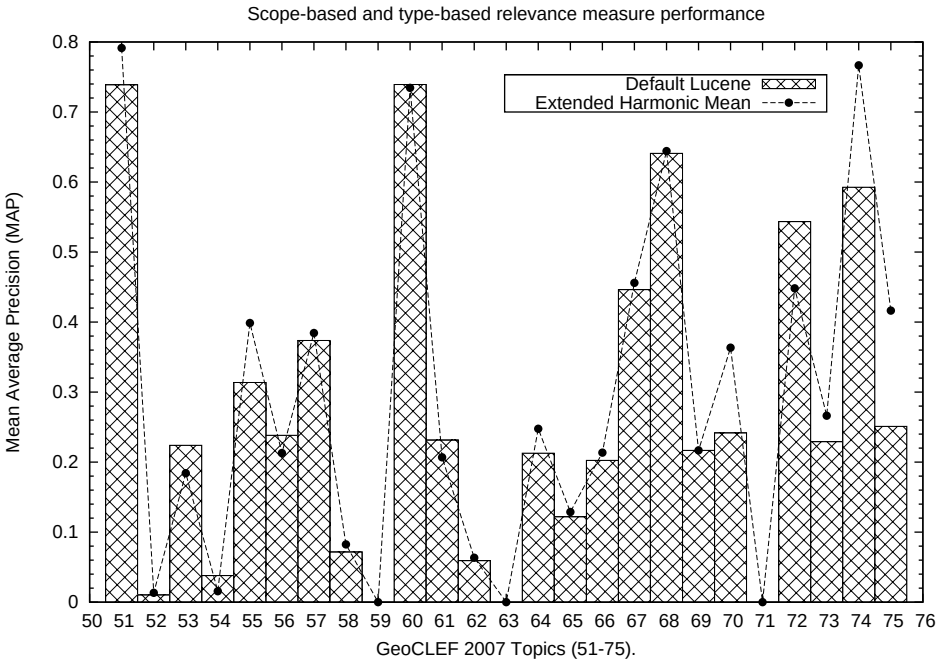

Figure 7.6: GeoCLEF 2007 per topic performance.

7.6 Conclusion

This chapter described two relevance ranking schemes that exploit geographical scopes and feature types in documents and search queries to rank documents by geography. The scope-based metric is used to rank documents for queries which are resolvable to at least one scope. On the other hand, the type-based metric is used to rank documents when a query mentions only the geographical subjects, so that it is not resolvable to a scope. The research objective addressed in this chapter is – How well can geographical scope and feature type information be integrated into the document ranking procedure to prioritize documents by geography?

The non-geographic and geographic relevance scores are combined through a linear interpolation and weighted harmonic-means. The harmonic mean-based combination achieved a better performance than linear interpolation. A better performance is achieved when more importance is attached to non-geographical retrieval than geographical retrieval. The best performance is achieved with harmonic mean derived formula with MAP scope of *0.2935*, an *8.9%* improvement over standard search engine.

Although the performance is encouraging, more research is needed to best incorporate geographical information mined from documents and search queries into the relevance ranking algorithm to improve search engine performance when answering geographically constrained information needs.